Data Mining in Earth System Science (DMESS 2015)

Forrest M. Hoffman[1,2], Jitendra Kumar[1], J. Walter Larson[3]

[1]Oak Ridge National Laboratory, [2]University of California–Irvine, and [3]Australian National University

June 2, 2015

International Conference on Computational Science (ICCS 2015)
Reykjavík University, Reykjavík, Iceland

Introduction

- Earth science data span many orders of magnitude in space and time scales.
- These data are increasingly large and complex, often representing long time series, making them difficult to analyze, visualize, interpret, and understand.
- Electronic data storage and high performance computing capacity enable creation of large data repositories and detailed empirical and process-based models.
- The resulting "explosion" of heterogeneous, multi-disciplinary Earth science data requires use of new analysis methods and development of highly scalable software tools.

Earth Science Data

- Observational and modeled data encompass temporal scales of seconds to millions of years (10^0–10^{13} s) and spatial scales of microns to tens of thousands of kilometers (10^{-6}–10^7 m).
- Integrating and synthesizing data across Earth science disciplines offers new opportunities for scientific discovery.
- The rise of data-centric science is becoming recognized as the *fourth paradigm of discovery* alongside the experimental, theoretical, and computational archetypes (Hey et al., 2009).
- However, the promise of data-intensive Earth science has yet to be realized because of the unique technological and social challenges it poses.

Model Results

- Open and user-friendly access to Earth science data is required—particularly for climate science—as interest in sustainability and environmental policy has added decision-makers and the public to the list of data users.
- Organized global climate modeling activities, like the Coupled Model Intercomparison Project (CMIP), can generate tens of terabytes to several petabytes of simulation results (Overpeck et al., 2011).
- CMIP results are now made available to the research community and the public through distributed, interconnected servers called the Earth System Grid (ESG; Williams et al., 2009).
- Composited, summary data from collections of simulation output are being developed to make model results more directly useful outside of the climate science community.

Observational Data

- Satellite remote sensing data tend to be very large and grow quickly as spatial and temporal resolutions increase.
- Meanwhile, small ecological data sets are often the most valuable for synthesis, but may be the hardest to preserve, distribute, and use (Reichman et al., 2011).
- Data curation and provenance must be formally documented; data format standards and metadata conventions are needed.
- Scientific workflow systems are being developed to document and automate data processing, quality control, gap-filling, analysis, and synthesis.
- The DataONE project (`http://www.dataone.org/`) is pioneering technologies to automate and document every step, from data acquisition and generation to synthesis and publication.

Model Validation Using Measurements

- Model evaluation places new demands on the measurements community to provide observations and uncertainties useful for assessing model fidelity (Randerson et al., 2009).
- Researchers need agreed-upon standards for benchmarks of scientific model performance.
- The International Land Model Benchmarking (ILAMB) project (`http://www.ilamb.org/`) was recently established to develop benchmarks for terrestrial biogeochemistry models.
- ILAMB will create a reusable and extensible, open source framework for evaluating metrics and generating diagnostics.
- By using freely available observational data and distributing its evaluation tools, ILAMB seeks to achieve a new standard for scientific openness and transparency (Kleiner, 2011).

Data Mining Approaches

- Much of today's large and complex Earth science data cannot be synthesized and analyzed using traditional methods on small desktop computers.
- Data mining algorithms and tools can be used to extract knowledge and information from observations and model data.
- Data mining, machine learning, and high performance visualization approaches that exploit distributed-memory parallel computational resources offer promising alternatives.
- Techniques include:
  - complex object-based image analysis (COBIA),
  - generalized extreme value (GEV) distributions,
  - support vector machines (SVMs),
  - self-organized maps (SOMs), and
  - cluster analysis.

Presentations

- *Pattern-Based Regionalization of Large Geospatial Datasets Using COBIA* – **Tomasz Stepinski**, Jacek Niesterowicz, Jaroslaw Jasiewicz
- *Fidelity of Precipitation Extremes in High Resolution Global Climate Simulations* – **Salil Mahajan**, Katherine Evans, Marcia Branstetter, Valentine Anantharaj, Juliann Leifeld
- *On Parallel and Scalable Classification and Clustering Techniques for Earth Science Datasets* – Markus Götz, Matthias Richerzhagen, Gabriele Cavallaro, Christian Bodenstein, Philipp Glock, **Morris Riedel**, Jon Atli Benediktsson
- *Completion of a sparse GLIDER database using multi-iterative Self-Organizing Maps (ITCOMP SOM)* – **Anastase - Alexander Charantonis**, Pierre Testor, Laurent Mortier, Fabrizio D'Ortenzio, Sylvie Thiria
- *A Feature-first Approach to Clustering for Highlighting Regions of Interest in Scientific Data* – **Robert Sisneros**

Program Committee

Thanks to our Program Committee for thier hard work:

- Michael W. Berry (University of Tennessee, USA)
- Bjørn-Gustaf J. Brooks (USDA Forest Service, USA)
- Nathaniel O. Collier (Oak Ridge National Laboratory, USA)
- Auroop R. Ganguly (Northeastern University, USA)
- William W. Hargrove (USDA Forest Service, USA)
- Forrest M. Hoffman (Oak Ridge National Laboratory, USA)
- Jian Huang (University of Tennessee, USA)
- Jitendra Kumar (Oak Ridge National Laboratory, USA)
- Vipin Kumar (University of Minnesota, USA)
- J. Walter Larson (The Australian National University, AUSTRALIA)
- Miguel D. Mahecha (Max Planck Institute for Biogeochemistry, GERMANY)
- Kumar Mahinthakumar (North Carolina State University, USA)
- Richard T. Mills (Intel Corp., USA)
- Stephen P. Norman (USDA Forest Service, USA)
- Karsten Steinhaeuser (University of Minnesota, USA)
- R. Raju Vatsavai (North Carolina State University, USA)
- Min Xu (Oak Ridge National Laboratory, USA)

References

W. W. Hargrove, F. M. Hoffman, and P. F. Hessburg. Mapcurves: A quantitative method for comparing categorical maps. *J. Geograph. Syst.*, 8(2):187–208, July 2006. doi: 10.1007/s10109-006-0025-x.

T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Corporation, Redmond, Washington, USA, Oct. 2009. ISBN 978-0-9825442-0-4.

F. M. Hoffman. Analysis of reflected spectral signatures and detection of geophysical disturbance using hyperspectral imagery. Master's thesis, University of Tennessee, Department of Physics and Astronomy, Knoxville, Tennessee, USA, Nov. 2004.

K. Kleiner. Data on demand. *Nature Clim. Change*, 1(1):10–12, Apr. 2011. doi: 10.1038/nclimate1057.

J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling. Climate data challenges in the 21st century. *Science*, 331(6018):700–702, Feb. 2011. doi: 10.1126/science.1197869.

J. T. Randerson, F. M. Hoffman, P. E. Thornton, N. M. Mahowald, K. Lindsay, Y.-H. Lee, C. D. Nevison, S. C. Doney, G. Bonan, R. Stöckli, C. Covey, S. W. Running, and I. Y. Fung. Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Global Change Biol.*, 15(10):2462–2484, Oct. 2009. ISSN 1365-2486. doi: 10.1111/j.1365-2486.2009.01912.x.

O. J. Reichman, M. B. Jones, and M. P. Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018):703–705, Feb. 2011. doi: 10.1126/science.1197962.

M. A. White, F. Hoffman, W. W. Hargrove, and R. R. Nemani. A global framework for monitoring phenological responses to climate change. *Geophys. Res. Lett.*, 32(4):L04705, Feb. 2005. doi: 10.1029/2004GL021961.

D. N. Williams, R. Drach, R. Ananthakrishnan, I. T. Foster, D. Fraser, F. Siebenlist, D. E. Bernholdt, M. Chen, J. Schwidder, S. Bharathi, A. L. Chervenak, R. Schuler, M. Su, D. Brown, L. Cinquini, P. Fox, J. Garcia, D. E. Middleton, W. G. Strand, N. Wilhelmi, S. Hankin, R. Schweitzer, P. Jones, A. Shoshani, and A. Sim. The Earth System Grid: Enabling access to multimodel climate simulation data. *Bull. Am. Meteorol. Soc.*, 90(2): 195–205, Feb. 2009. doi: 10.1175/2008BAMS2459.1.

Acknowledgements

U.S. DEPARTMENT OF ENERGY

Office of Science

This research was sponsored by the Climate and Environmental Sciences Division (CESD) of the Biological and Environmental Research (BER) Program in the U. S. Department of Energy Office of Science, the U. S. Department of Agriculture Forest Service, and the National Science Foundation (AGS-1048890). This research used resources of the National Center for Computational Sciences (NCCS) at Oak Ridge National Laboratory (ORNL), which is managed by UT-Battelle, LLC, for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.